Published online 4 June 2012 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* ***W329–W333***
doi:10.1093/nar/gks488

CMWeb: an interactive on-line tool for analysing residue–residue contacts and contact prediction methods

Dániel Kozma, István Simon and Gábor E. Tusnády*

Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, PO Box 7, H-1518 Budapest, Hungary

Received February 24, 2012; Revised April 27, 2012; Accepted May 3, 2012

ABSTRACT

A contact map is a 2D derivative of the 3D structure of proteins, containing various residue–residue (RR) contacts within the structure. Contact maps can be used for the reconstruction of structure with high accuracy and can be predicted from the amino acid sequence. Therefore understanding the various properties of contact maps is an important step in protein structure prediction. For investigating basic properties of contact formation and contact clusters we set up an integrated system called Contact Map Web Viewer, or CMWeb for short. The server can be used to visualize contact maps, to link contacts and to show them both in 3D structures and in multiple sequence alignments and to calculate various statistics on contacts. Moreover, we have implemented five contact prediction methods in the CMWeb server to visualize the predicted and real RR contacts in one contact map. The results of other RR contact prediction methods can be uploaded as a benchmark test onto the server as well. All of these functionality is behind a web server, thus for using our application only a Java-capable web browser is needed, no further program installation is required. The CMWeb is freely accessible at http://cmweb.enzim.hu.

INTRODUCTION

Structures of globular proteins are determined and maintained by non-covalent residue residue–(RR) interactions (1). Mapping RR contacts into a 2D binary map results in the so called contact map. Contact maps can be predicted from amino acid sequence information of proteins with acceptable accuracy. Several methods have been developed for predicting these contacts based on machine learning algorithms (2–4), or simpler statistical-based algorithm like mutual information (MI) (5), correlated mutations (6,7) and statistical coupling algorithm (SCA) (8), etc. The accuracy of the state-of-the-art RR predictors is ~20–30%, suggesting the need for improvement, although the most recent methods [e.g. (9)] show significantly better, but still unsatisfactory performance.

To understand the properties of contact maps and the relations between of 3D structure and residue contacts, besides statistical approaches, visual inspection of contact maps can be useful, as well. During the last decades several useful contact map viewers have been developed (10–12). The most recent contact map viewers is the CMView (13) program using PyMol (14) for visualizing the 3D structures. CMView is a desktop application, which is mainly designed for studying 3D structure reconstruction from a contact map. The Contact Map Web Viewer (CMWeb) server presented in this article has a different purpose. CMWeb is designed for analysing, understanding contact formation, protein contacts and to help to develop methods for predicting protein contacts. Our aim is not predicting 3D structure of proteins, only the visual investigation of RR contacts and the results of RR contact prediction methods.

The server is a standalone, user-friendly platform, which does not require additional component for operation.

MATERIALS AND METHODS

The web server is written in C/C++ using the Wt web toolkit (15) and the in house written PDBLIB program library used for TMDET algorithm earlier (16). In contempt of the numerous calculations the web server is really fast due to the C/C++ program core. The web server utilizes the OpenAstex (17) protein structure viewer for combined contact map and 3D structure

*To whom correspondence should be addressed. Tel: +361 2793159; Fax: +361 4665465; Email: tusnady.gabor@ttk.mta.hu

view. We choose OpenAstex structure viewer because it renders molecules more nicely and faster than other such Java-based methods. We suggest to utilize at the client browser the most commonly used Oracle Java JRE (http://java.com).

As the web server is designed for analysing protein chains and structures, we apply a basic filter on PDB entries to exclude nucleic acid structures.

Multiple sequence alignment

The multiple sequence alignment (MSA) is generated, based on the sequence stored in the PDB file. The sequence is searched against a user selectable sequence database (SwissProt or nr) using the BLAST algorithm. MSA is generated from the resulted local pair alignments, where the columns containing gap in the query sequence are neglected. The prediction methods use this generated MSA for estimating contacts.

Implemented contact prediction methods

Five protein contact prediction methods have been implemented as follows: MI (Mutual Information) (5), SCA (Statistical Coupling Analysis) (8), ELSC (Explicit Likelihood of Subset Co-variation) (7), OMES (Observer Minus Expected Squared) (18) and the one of the first methods by Göbel (6). We have re-implemented these contact predictors in C programming language to make on-the-fly prediction realizable. Because our aim is to benchmark these or any other methods, predictions can be made only using PDB entries, user provided sequences are not allowed. These implementations were checked on the original as well as on other tested implementation of these prediction methods.

RESULTS AND DISCUSSION

Contact map viewer

The CMWeb server integrates a contact map, structure and MSA viewer combined with a statistical evaluating system (Figure 1) in a fully interactive way. The web server provides a graphical user interface (GUI) like web application, where the various objects on the screen are connected via signals, therefore any user interaction is traced and handled by these objects. When selecting a contact pair in the contact map panel, the server executes the following processes:

(i) shows the corresponding residues in the structure viewer panel (Figure 1D), coloured corresponding to the secondary structure scheme shown in the MSA panel (Figure 1F), where conservation profile can be found as well;
(ii) highlights positions in the sequence alignment (Figure 1F);
(iii) displays the distance between the selected residues in the structure viewer proportional to the contact definition (Figure 1D);
(iv) displays the distance value, the residue number and the connecting atom types in the information panel (Figure 1C).

Furthermore, the web server could shows all the neighbours of a selected residue (using double-click on the contact map panel) in the structure viewer. The central residue and its neighbours are coloured by a given colour scheme. The centre of the given cluster and the number of the surrounding residues are displayed in the information panel. The position of the selected residues are highlighted in the sequence alignment panel as well. These functions can be also activated from the MSA panel. Additionally selecting any region in the MSA is displayed in the structure viewer too.

All data presented on the webpage are calculated on-the-fly based on the selected or uploaded PDB protein chain structure. The contact definition can be specified in terms of contact type (all-atom, side-chain atoms, C_α, and C_β) and contact threshold (distance cutoff in Å). In addition, the user can filter indirect contacts within a given distance limit. Furthermore, the contact map could display the indirect connections of the residues over heteromolecules such as e.g. structural waters. The server can shows the contact map proportional to the contact definition (Figure 2A) or the distance matrix of a given protein chain (Figure 2B).

The user can investigate all PDB entry by entering the PDB code or can upload any protein structure in PDB format for visualizing own, not published or modelled structures, as well. The server incorporates all PDB entries and is updated weekly.

The server provides MSA with schematic secondary structure and conservation profile as well, to help us to collect necessary sequence information, whereas similar sequences share roughly the same structure.

The main advantage of these features is that we can get broad information with one click about the inspected residues and its physco-chemical, spatial environment with highlighting and displaying the corresponding positions in the MSA panel and in the structure viewer simultaneously.

The web server calculates statistics on the inspected protein chain. A marginal and a double marginal distribution of amino acid contact numbers are presented. The later shows the population of RR contacts between amino acids with n and m number of contacts, various amino acid frequencies and RR contact distribution. The predicted results displayed on the contact map panel and the performance of the given method on the specified protein chain is shown by the ROC curve. In addition we can follow with attention the separation of the TP or FP scores, and the informative statistical measures such as accuracy, precision, TPR/sesitivity/coverage/recall, FPR, Matthews correlation coefficient, improvement over random, F1 and Xd scores. In addition to a ROC curve, score histogram and statistical measures for evaluating performance of prediction techniques are presented, as well. The score histogram is a useful check of the prediction methods, here we can see the separation of the score values calculated for residue pairs are in contact and for which are not.

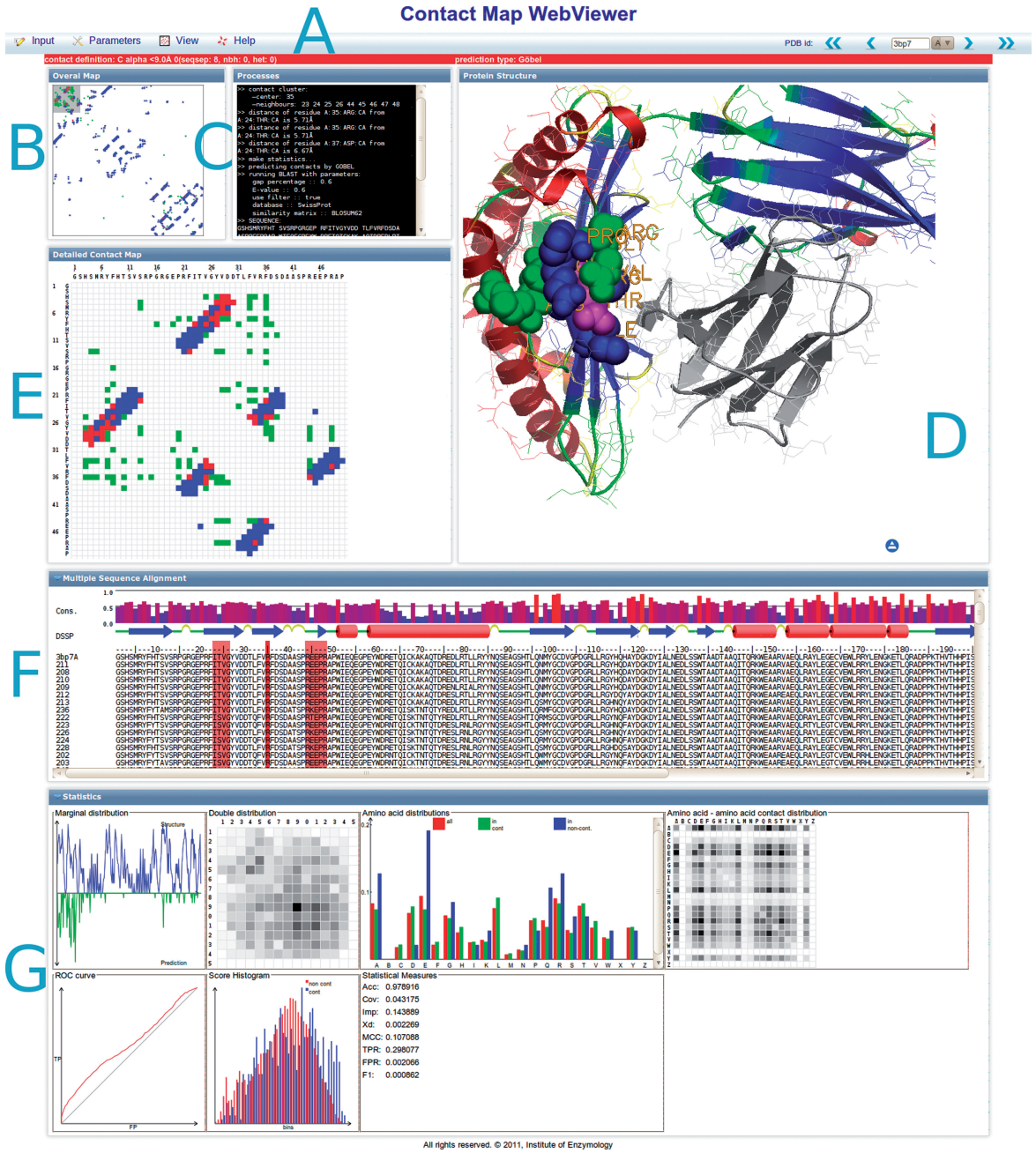

Figure 1. Layout of the CMWeb web server. (**A**) menu bar and navigation bar; (**B**) overall contact map; (**C**) information panel; (**D**) structure viewer; (**E**) zoomable detailed contact map (blue: contacts, green: false prediction, red: correct prediction); (**F**) MSA viewer with conservation profile and secondary structure; (**G**) statistical panel with marginal and double marginal distribution of contacts, amino acid distributions, amino acid contact propensities, ROC curve, score histogram and a table of statistical measures.

Benchmark test

Using the benchmark test menu users can check the performance of any contact prediction methods. After uploading the CASP RR formatted prediction files and setting contact map definition our statistical evaluating system returns a list with the name, small contact map including predictions and different statistical measures line by line. Each prediction can be analysed further (Figure 3) inspecting the contacts between residues and orientation of them in the 3D space using the OpenAstex(17) molecular viewer described above. The summary at the end of the list gives a brief information about the average performance of the tested contact prediction technique. It is important to note that our evaluating system neglect the distance ranges in the file, it evaluates predictions based on the contact map definition previously set by the user.

CONCLUSION

CMWeb is an interactive on-line web application to examine contact maps together with linked 3D structures,

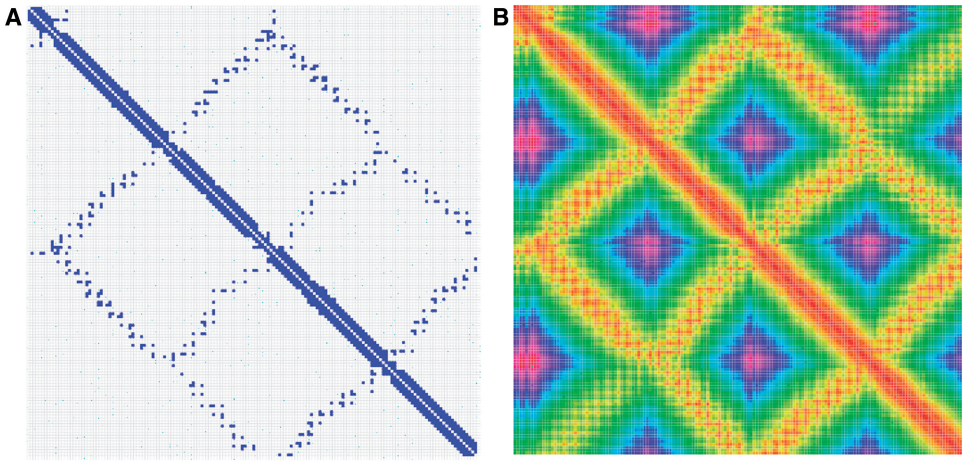

Figure 2. Contact maps of the 2bl2A protein chain with the given contact definition. (**A**) Binary contact map (any heavy atoms closer than the sum of their van der Waals radii plus 1.5Å, the sequence separation is 1); (**B**) Continuous distance map (distance scale is from the closest to the farthest as red–yellow–green–blue–purple).

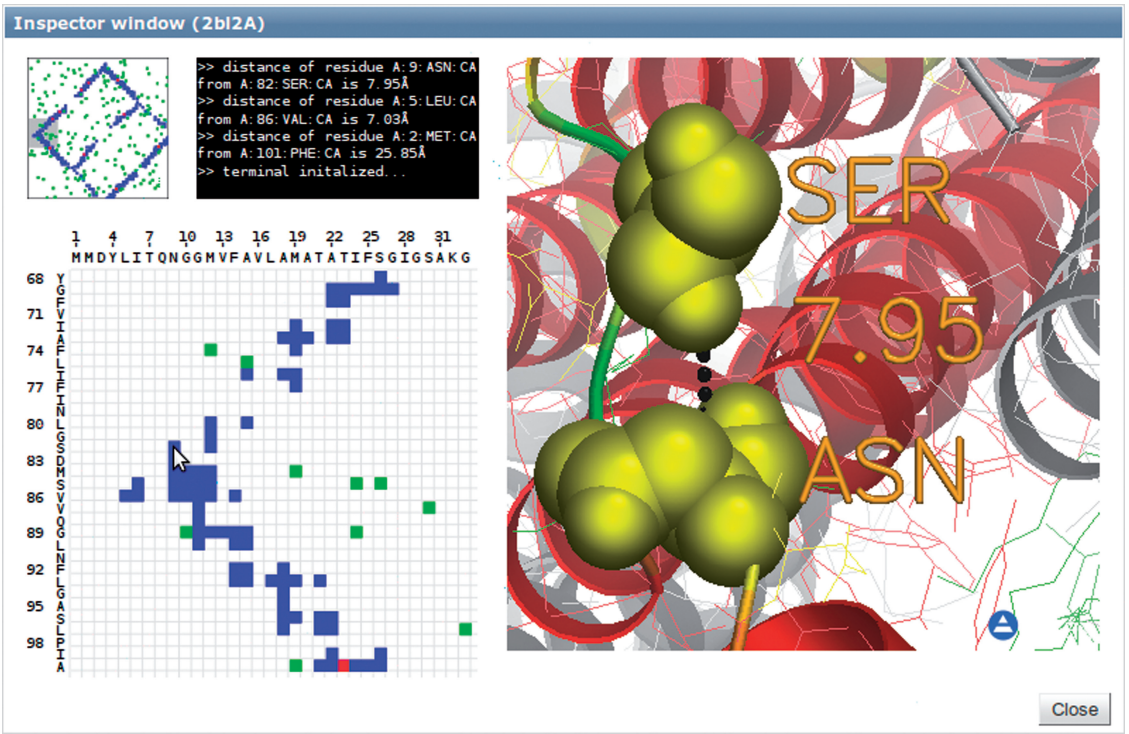

Figure 3. Layout of the inspector window which is useful for the further analysis of the elements of the benchmark setlist, we could inspect the environment of the correct and incorrect predictions using the structure viewer on the right. In the upper left corner there is an overall contact map and an information box displaying the user activity, in the bottom left corner there is a detailed contact map, the blue points are the real contacts corresponding to the contact definition, the reds are the correct and greens are the false predictions.

MSAs, secondary structures, sequence conservation and five commonly used prediction methods. Furthermore, CMWeb can be used for benchmark testing custom prediction methods and measuring theirs performance. The server utilize state-of-the-art technologies to provide a desktop application like GUI and functionality on the web. This web server could be a good example of the hidden great potential of the Wt programming library. We hope CMWeb will be a powerful web tool for analysing protein contacts and contact prediction methods and may become a widespread scientific tool.

ACKNOWLEDGEMENTS

Comments on the manuscript by Mónika Fuxreiter and on the manual of the CMWeb server by Bálint Mészáros

are gratefully acknowledged. Finally we would like to acknowledge for the help from Koen Deforche in the development of CMWeb.

FUNDING

Hungarian Scientific Research Fund (OTKA) [NK100482 and K75460]. Funding for open access charge: Research grant of Hungarian Scientific Research Fund.

Conflict of interest statement. None declared.

REFERENCES

1. Michael Gromiha,M. and Selvaraj,S. (2004) Inter-residue interactions in protein folding and stability. *Prog. biophy. Mol. Biol.*, **86**, 235–277.
2. Fariselli,P. and Casadio,R. (1999) A neural network based predictor of residue contacts in proteins. *Protein Eng.*, **12**, 15–21.
3. Punta,M. and Rost,B. (2005) PROFcon: novel prediction of long-range contacts. *Bioinformatics*, **21**, 2960–2968.
4. Xue,B., Faraggi,E. and Zhou,Y. (2009) Predicting residue-residue contact maps by a two-layer, integrated neural-network method. *Proteins.*, **76**, 176–183.
5. Atchley,W.R., Wollenberg,K.R., Fitch,W.M., Terhalle,W. and Dress,A.W. (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol.*, **17**, 164–178.
6. Göbel,U., Sander,C., Schneider,R. and Valencia,A. (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.
7. Dekker,J.P., Fodor,A., Aldrich,R.W. and Yellen,G. (2004) A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, **20**, 1565–1572.
8. Lockless,S.W. and Ranganathan,R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
9. Jones,D.T., Buchan,D.W.A., Cozzetto,D. and Pontil,M. (2011) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.
10. Sonnhammer,E.L. and Wootton,J.C. (1998) Dynamic contact maps of protein structures. *J. Mol. Graph. Model.*, **16**, 1–5, 33.
11. Chung,J.L., Beaver,J.E., Scheeff,E.D. and Bourne,P.E. (2007) Con-Struct map: a comparative contact map analysis tool. *Bioinformatics*, **23**, 2491–2492.
12. Pietal,M.J., Tuszynska,I. and Bujnicki,J.M. (2007) PROTMAP2D: visualization, comparison and analysis of 2D maps of protein structure. *Bioinformatics*, **23**, 1429–1430.
13. Vehlow,C., Stehr,H., Winkelmann,M., Duarte,J.M., Petzold,L., Dinse,J. and Lappe,M. (2011) CMView: interactive contact map visualization and analysis. *Bioinformatics*, **27**, 1573–1574.
14. DeLano,W.L. (2002) *Pymol Molecular Graphics System.*, http://www.pymol.org (2012, date last accessed).
15. Deforche,K. (2009) *Wt a C++ Web Toolkit*, http://www.webtoolkit.eu/wt (2012, date last accessed).
16. Tusnády,G.E., Dosztányi,Z. and Simon,I. (2004) Transmembrane proteins in the protein data bank: identification and classification. *Bioinformatics*, **20**, 2964–2972.
17. Hartshorn,M.J. (2002) AstexViewer: a visualisation aid for structure-based drug design. *J. Comput. Aid. Mol. Des.*, **16**, 871–881.
18. Kass,I. and Horovitz,A. (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins*, **48**, 611–617.